AD___

Award Number: W81XWH-12-1-0298

TITLE: MTHFR Functional Polymorphism C677T and Genomic Instability in the Etiology of Idiopathic Autism in Simplex Families

PRINCIPAL INVESTIGATOR: Xudong Liu, PhD

CONTRACTING ORGANIZATION: Queen's University
Kingston, K7M 8A6 Canada

REPORT DATE: October 2013

TYPE OF REPORT: Revised Annual Report

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
October 2013	Revised Annual Report	30September2012-29September2013

4. TITLE AND SUBTITLE
MTHFR Functional Polymorphism C677T and Genomic Instability in the Etiology of Idiopathic Autism in Simplex Families

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-12-1-0298

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Xudong Liu, PhD

email: liux@queensu.ca

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Queen's University
Kingston, K7M 8A6 Canada

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
With our previous published study revealed significantly association of C677T polymorphism in MTHFR gene with idiopathic autism in Simplex (SPX) autism families; and the proven facts that de novo CNVs rates are consistently high in SPX ASD (5.8%-10.2%) versus familial ASD (2-3%), we hypothesize that low-activity MTHFR 677T allele leads to increase global DNA hypomethylation and consequently results in increased generation of de novo CNVs bringing about a higher risk for developing sporadic cases of autism. We proposed to test 1) the association of MTHFR 677T allele with rate of ASD related de novo CNVs; 2) the association of of MTHFR 677T allele with increased level of global hypomethylation; and 3) the association of level of global hypomethylation with increased rate of ASD related de novo CNVs.

15. SUBJECT TERMS- none provided

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	14	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

Page

INTRODUCTION: With our previous published study revealed significantly association of C677T polymorphism in MTHFR gene with idiopathic autism in Simplex (SPX) autism families; and the proven facts that *de novo* CNVs rates are consistently high in SPX ASD (5.8%-10.2%) versus familial ASD (2-3%), we hypothesize that low-activity MTHFR 677T allele leads to increase global DNA hypomethylation and consequently results in increased generation of de novo CNVs bringing about a higher risk for developing sporadic cases of autism. We proposed to test 1) the association of MTHFR 677T allele with rate of ASD related *de novo* CNVs; 2) the association of of MTHFR 677T allele with increased level of global hypomethylation; and 3) the association of level of global hypomethylation with increased rate of ASD related *de novo* CNVs.

BODY: This pilot project started from September of 2013. Although no publication has yet come out from the study during this first year period, we did make significant progress to reach the milestones set in our proposal.

We have reviewed and analyzed the microarray data from Affymetrix 2.7M Cyto array or Affymetrix SNP6.0 platforms or other high resolution array platforms on autism cases from 5100 SPX families, with another 10 families still to be done by the end of February of 2014. CNVs were called as proposed in the grant. 99 individuals were found to carry pathogenic CNVs, which was confirmed by real-time quantitative PCR. 33 of these CNVs were determined to be de novo using available parents' array data or RTqPCR results. We anticipate to have another 10 cases with de novo pathogenic CNVs to be identified in the additional families.

All the available families (both parents and affected individuals) were genotyped on MTHFR Functional Polymorphism C677T using TaqMan Assay from Life Technologies.

A experimental pipeline was established for global Methylation profiling using Medip-Seq strategy with Ion Torrent Proton: Including optimal conditions for Methyly-Capturing, Sequencing Template Prep and Sequencing Depth and Coverage selection. A bioinformatics data analysis pipeline under Linux environment was also established to analyze the global methylation sequencing data for DMR, and comparison between groups of individuals.

With above pipelines, 15 ASD cases with de novo pathogenic CNVs and 15 ASD cases without de novo pathogenic CNVs (based on th results from above 1) and 2)) were Medip-Sequenced, the data analysis is on-going.

We anticipate completing all the experiment on time by February of 2014 and finish the data analysis by April of 2014 and move forward with manuscript preparation and final report.

KEY RESEARCH ACCOMPLISHMENTS:

- Completed microarray data analysis on 510 SPX families; identified 99 individuals carrying pathogenic CNVs; confirmed 33 of these CNVs be de novo.
- 510 SPX families (both parents and affected individuals) were genotyped on MTHFR Functional Polymorphism C677T using TaqMan Assay.
- Both experimental and bioinformatics pipelines were established for global Methylation profiling using Medip-Seq strategy with Ion Torrent Proton
- 15 ASD cases with de novo pathogenic CNVs and 15 ASD cases without de novo pathogenic CNVs were Medip-Sequenced, the data analysis is on-going.

REPORTABLE OUTCOMES: This pilot study is still ongoing and has not resulted in publication yet so far.

CONCLUSION: We anticipate to completing the project by April of 2014, and reporting the findings that may have potential significance. .

REFERENCES: N/A
APPENDICES: see attached
SUPPORTING DATA: see attached

Supplementary Document for Progress Summary

I. Genotyping of MTHFR C677T Polymorphism

The functional SNP in the MTHFR gene, C677T (rs1801133), was genotyped using validated custom TaqMan SNP Genotyping Assays (http://www.appliedbiosystems.com) on an ABI Prism 7900HT in 510 families so far. Genotypes were automatically scored with the SDS 2.2.2 software using standard parameters. Allele and genotype distributions of MTHFR C677T in children with autism and their parents were presented in Table 1.

Table 1: Allele and genotype distributions of MTHFR C677T in children with autism and their parents

	Allele Distribution			Genotype Distribution			
	N	**C**	**T**		**C/C**	**C/T**	**T/T**
Affected Children	1020	579 (56.8%)	441 (43.2%)	510	167 (32.7%)	245 (48.0%)	98 (19.2%)
Parents	2040	1314 (64.4%)	726 (35.6%)	1020	429 (42.1%)	456 (44.7%)	135 (13.2%)

II. Global Methylation Profiling Using MBD-NGS Seq(Methylated DNA enrichment and Next Generation Sequencing)

Genomic DNA from ASD cases is fragmented using enzymatic methods with Ion Xpress™ Plus Fragment Library Kit (http://www.lifetechnologies.com/order/catalog/product/4471269). Optimized conditions resulted in DNA fragments ranging from 50b to200 bp in length. The fragmented DNA was then end repaired and ligated to IonXpresss library adaptors. Fragmented DNA library was then subjected to MethylMinerTM methylated DNA kit enrichment according to the manufacturer's protocol eluted the methylated DNA as a single fraction with buffer containing 2M NaCl. The enriched and captured library was then subjected to sequencing with Ion Torrent Pronto using PI chip according to standard protocol (https://www.lifetechnologies.com). Barcoding strategy was adopted to sequence multiple samples on one sequencing run

A summary of a sequencing run for 5 barcoded samples was presented below (attached at the end of this summary). So far 34 samples have been sequenced, the sequencing of remaining samples will be completed by April of 2014.

III. Methylome Data Analysis

We established a pipeline based on public available tools developed by other researchers to analyze the methylome sequencing data generated from **MBD-NGS Seq.** The pipeline was tested on a small number of samples sequenced so far. A complete analysis will be performed once all the samples are sequenced.

The analysis pipeline takes as input the raw methylome DNA sequences aligned to the human reference genome in the form of a bam files which were generated by primary analysis on Ion Proton Server after the sequencing run. Pre-processing of the files is done using ***samtools*** to remove duplicated reads and sort the bam files for use downstream. The pipeline can comparatively analyzes two sample groups using the R-

package MEDIPS. The package was designed specifically for the use of methylome sequencing data and determines the relative methylation score of each region along the genome. It works off of the principle that regions with greater methylation would be pulled down to a greater extent. Therefore the program looks for regions of overlapping 'CG' motifs within the bam file, assigning an expression value based on the number of overlaps. After adjusting for multiple samples testing the averaged values for each region of each group are used to determine if that region is differentially expressed.

The initial output of the pipeline focuses on quality control of the samples. Three different parameters are analyzed based on metrics provided by us to select only bam files which pass the QC. The three parameters are defined as:

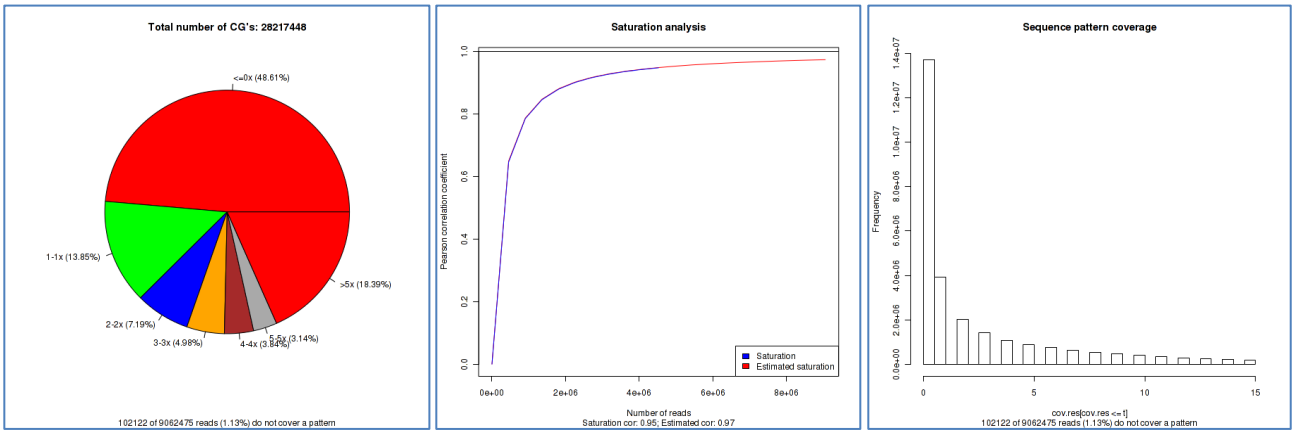
1. **Sequence Coverage:** The minimum percentage of reads which have 5x or > coverage within the bam file. This parameter is currently set at 5% however 5x coverage of most samples analyzed has been in excess of 20%
2. **CpG Enrichment:** for each given region the number of C's, G's, CpG's and total bases is used to determine an observed / expected ratio between the reference genome and the methylone being used. The regions pulled down using the methyl capture kits are expected to be enriched with methylation when compared to the genome, therefore to ensure that enriched reads were pulled down a minimum ratio of 1.7 is set. Non enriched DNA is expected to have a ration of around 1.
3. **Saturation Analysis:** The saturation analysis addresses the question, whether the given set of mapped reads is sufficient to generate a saturated and reproducible coverage profile of the reference genome. If there is a enough short reads the coverage profile can be reproduced by another independent set of similar short reads. The values for saturation range from 0-1 corresponding to the Pearson coefficient of correlation. The cut off employed in our pipeline is 0.5, however most samples produce a saturation score of around 0.9 indicating a highly saturated sample.

The pipeline automatically determines which bam files pass or fail the filters put in place, however at the same time it generates summary data, so that we may go back and visually analyze the quality control at a later point to potentially refine the QC parameters.

QC

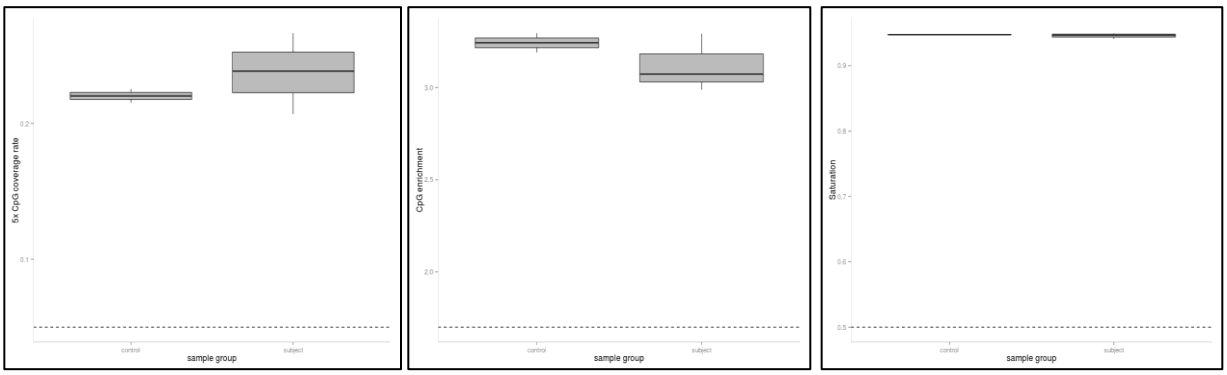
A standard QC produces the following outputs when run on a set of 5 bam files randomly assigned to either the control or the affected groups for testing (These samples are actual methylome data generated within our lab).
Coverage Analysis is represented as both a pie chart as well as a histogram. Eah image corresponds to the coverage of a single bam file.

The saturation Analysis displays both the true and the estimated saturation for one sample.



All of the graphs can easily be used by us to assess the quality of the data that was generated. For example, by looking at the histogram of sequence pattern coverage, it can be seen that there is a large amount of data at below 5x coverage for a given read and a given CG pattern, however greater than 20% of the data has at least 5x coverage. It is expected that there will be a significant amount of low coverage reads and 0x coverage CG's since a vast majority of the genome is not captured using methyl capture techniques. Therefore 20% coverage of 5x or greater is considered a pass of the QC. The high saturation score shows there is sufficient coverage across the genome for the results to be reproducible with a similar number of reads in another sample, therefore this file can be used for comparison.

Additionally, The CpG enrichment is given in a text file that can easily be read. The enrichment for this sample is 3.295 (observed / expected), indicating a region which is highly enriched in CpG's. Furthermore Summary boxplots are mad to compare the distribution of the parameters between all the samples to get a better idea of the data quality. The dotted line on each plot represents the Cut-off used for the QC.



Based on the above output, samples are either accepted or rejected for downstream analysis and determination of differential methylation.

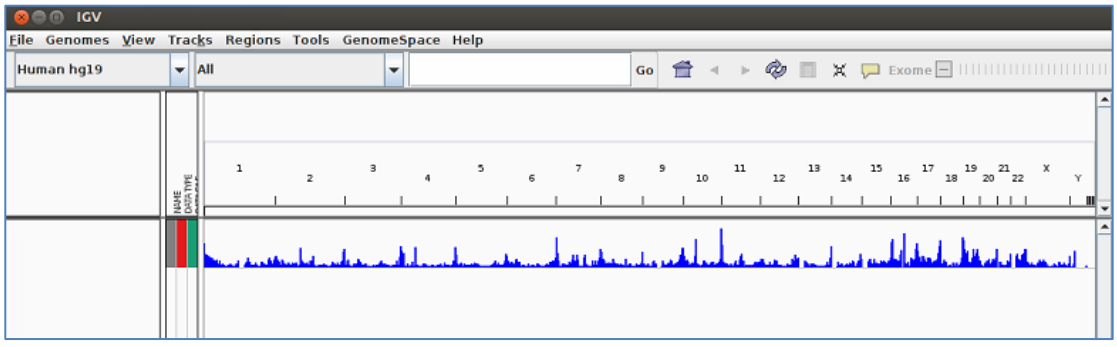
Discovery of differentially methylated regions Differentially Methylated Regions (DMR)

After QC DMR discovery can be completed. The process first normalizes the data using the reference genome converting the raw "counts" of CpG's within each 500bp window into a normalized score (Reads per Kilobase per Million). The RPKM values for each region are then compared and utilized to determine expression values for those regions (the control set is considered the normal expression). The expression values dictate what is considered differentially methylated, based on defined variables. For the data presented here regions which were differentially methylated had to have had an expression value of at least 2x greater than the control data set and an adjusted p-value of at least 0.01.

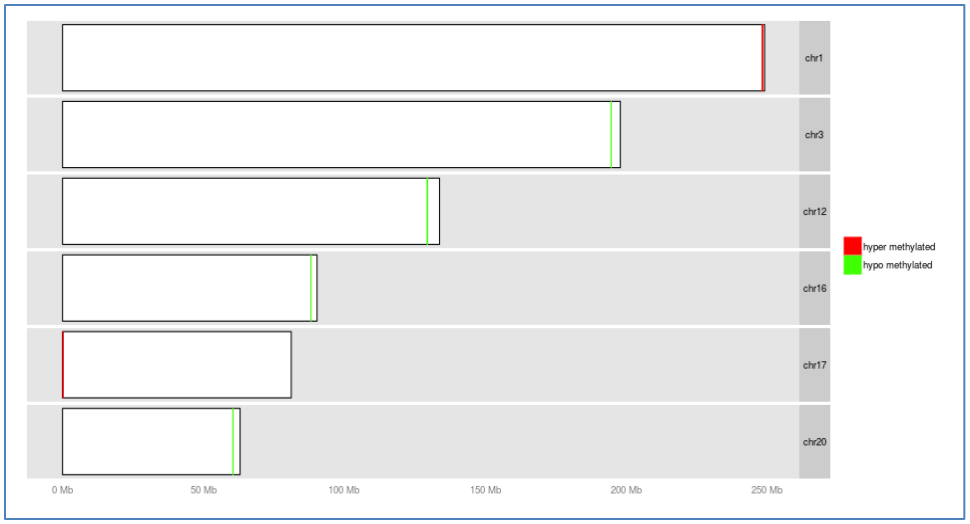
For statistical purposes the entire dataset generated is saved as a .Rdata file which can be readily loaded into R for further specific testing and analysis at a later time. The file contains all the relevant information in the form of a table which can easily be manipulated and interpreted. The table is annotated so that is contains the genes within each given window. Additional annotations can easily be added at any time for correlation with DMR's.

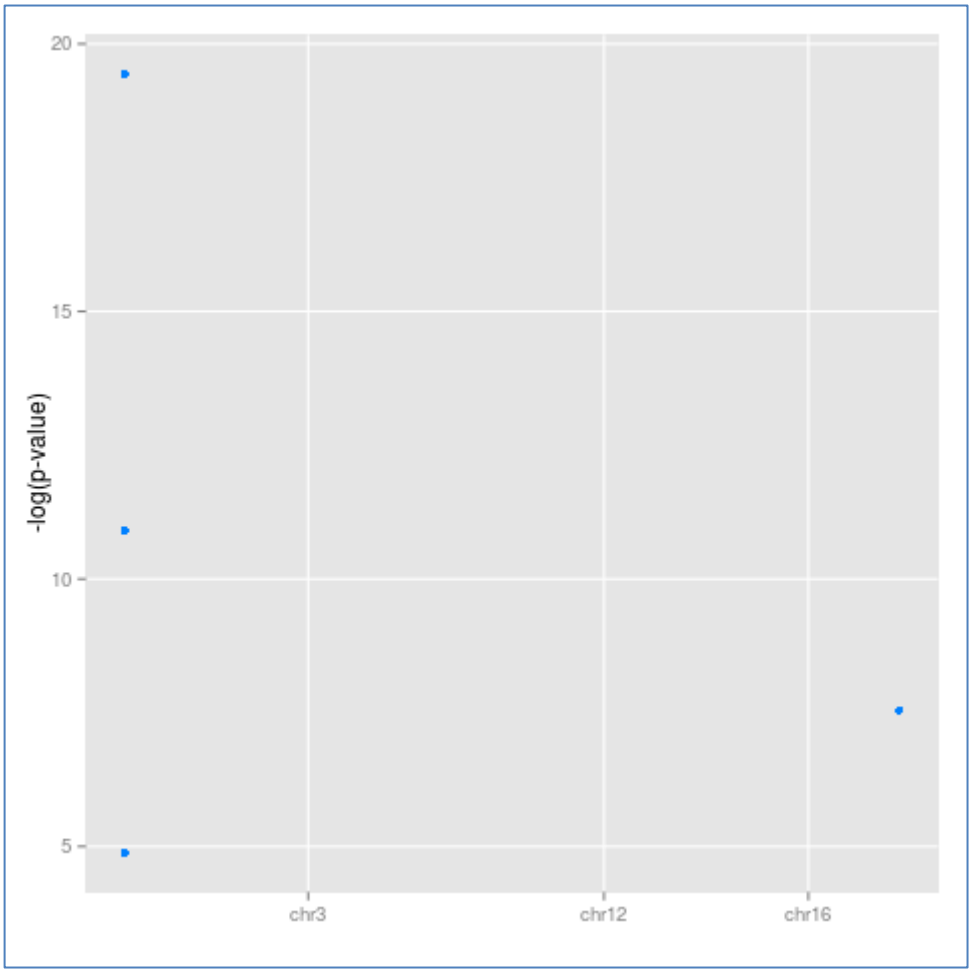
Once DMRs are identified, they are displayed in several methods. A tab delimited text file is produced containing all the information for each differentially methylated region, including the gene annotations which can easily be displayed in both R and a workbook such as excel. From this list further filter and annotation may be applied to find genes of interest which are differentially methylated for further investigation. From each bam file a wiggle (.wig) file is produced which can be used with various genome browsers. The wig file contains information on the RPKM values for all regions contained in the bam file. These values represent the relative methylation of each region and therefore serve as a visual reference of the methylated state of the sample.

Furthermore, Manhattan plots of the log10-Pvalues and Karyotype plots are produced to display the DMR's. These plots serve to visually represent the data generated. More information can easily be included in each graph depending on what we are specifically looking for. In this case, only the regions determined to be DMR's were displayed on their chromosomes.



An additional filter can be applied using annotation tools to only display regions which are intragenic, narrowing the search for relevant genes and DMRs. The samples which were ran were all from affected individuals however were randomly segregated into a case control for the purpose of displaying the capability of the pipeline. Because of this and potentially because of stringent settings few DMRs were detected, as seen by both the Manhattan plot and the karyotype.





The pipeline can also employ hierarchical cluster analysis algorithms to group either samples, or genes. This functionality can be changed based on the parameters of what we are searching for. It can be based solely on the DMRs or can also include the data from all genomic windows. This function is still currently under development.

The goal of this pipeline is twofold: to identify DMRs and potential genes which are differentially methylated in the affected individuals compared to the control, as well as assessing the global methylation score (GMS), determining whether there is significant differences between both the affected and the control. Actually scoring the state of methylation is currently under development. It is expected that this will be implemented in the near future to determine the average GMS of the affected vs. the control to determine if there are any significant difference between the two. The implementation of more functions within the pipeline to further analyze the data will be the next step in the development process, as well as inputting more samples for testing.

Future Steps

When a complete analysis is to to be carried out, we will be looking to implement additional QC metrics that may be used to increase the quality of the data. Furthermore we will work to include a GMS in order to test to see if there is any statistical difference between the methylation of the affected group and the control group. Finally, further refinement of the clustering algorithms, and the annotation and identification of specific gens will be important for the progression.

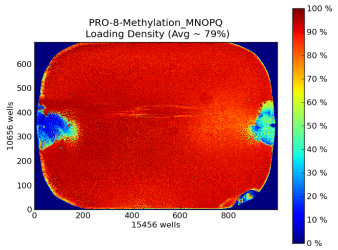
Attachment: a report of an ION Proton sequencing run of MDB captured methylated DNA with Ion PI chip for 5 barcoded samples.

Run Summary

5.6 G Total Bases | **59** Key Signal

79% ISP Loading

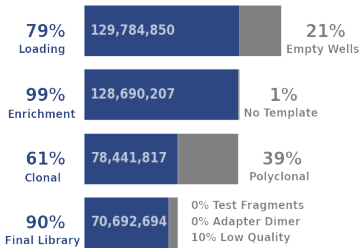
ISP Density



70,692,694 Total Reads

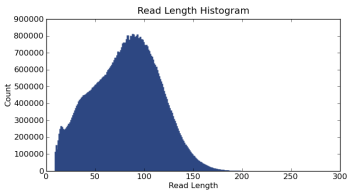
55% Usable Reads

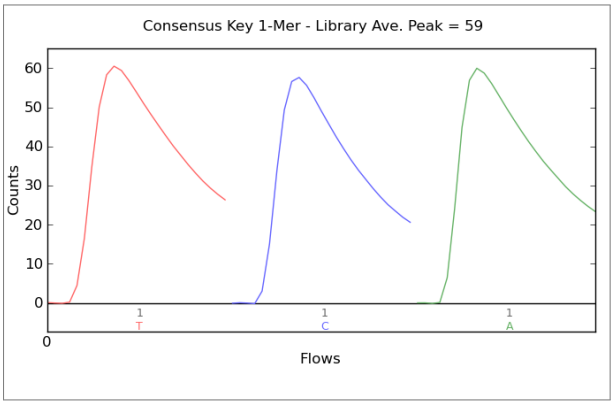
ISP Summary



80 bp Mean | **83 bp** Median | **88 bp** Mode

Read Length





Addressable Wells	**164,699,136**	
With ISPs	129,784,850	78.8%
Live	128,690,207	99.2%
Test Fragment	398,355	00.3%
Library	128,291,852	99.7%

Library ISPs	**128,291,852**	
Filtered: Polyclonal	50,248,390	39.2%
Filtered: Low Quality	7,262,039	05.7%
Filtered: Primer Dimer	88,729	00.1%
Final Library ISPs	**70,692,694**	**55.1%**

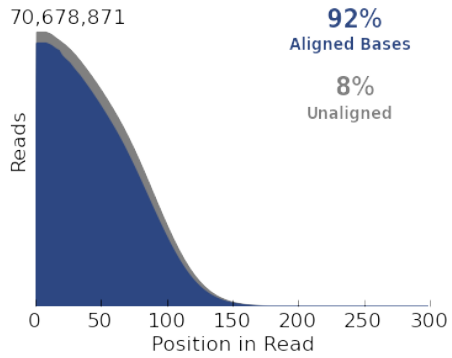
Barcode Name	Sample	Bases	$\geq Q20$	Reads	Mean Read Length
No barcode	None	41,471,743	27,880,370	590,610	70 bp
IonXpress_017	M-28294	1,165,059,311	830,529,788	14,432,612	80 bp
IonXpress_018	N-43592	1,329,337,126	967,648,773	16,388,826	81 bp
IonXpress_019	O-43364	1,042,704,966	763,253,222	13,046,004	79 bp
IonXpress_020	P-43614	989,379,437	714,281,294	12,373,034	79 bp
IonXpress_021	Q-43863	1,088,741,510	801,815,121	13,847,784	78 bp

Test Fragment	Reads	Percent 50AQ17	Read Length Histogram
TF_C	**287,587**	**18%**	

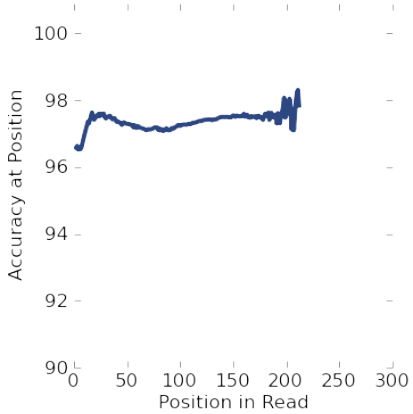
Alignment Summary *(aligned to hg19)*

5.1 G
Total Alignment Bases

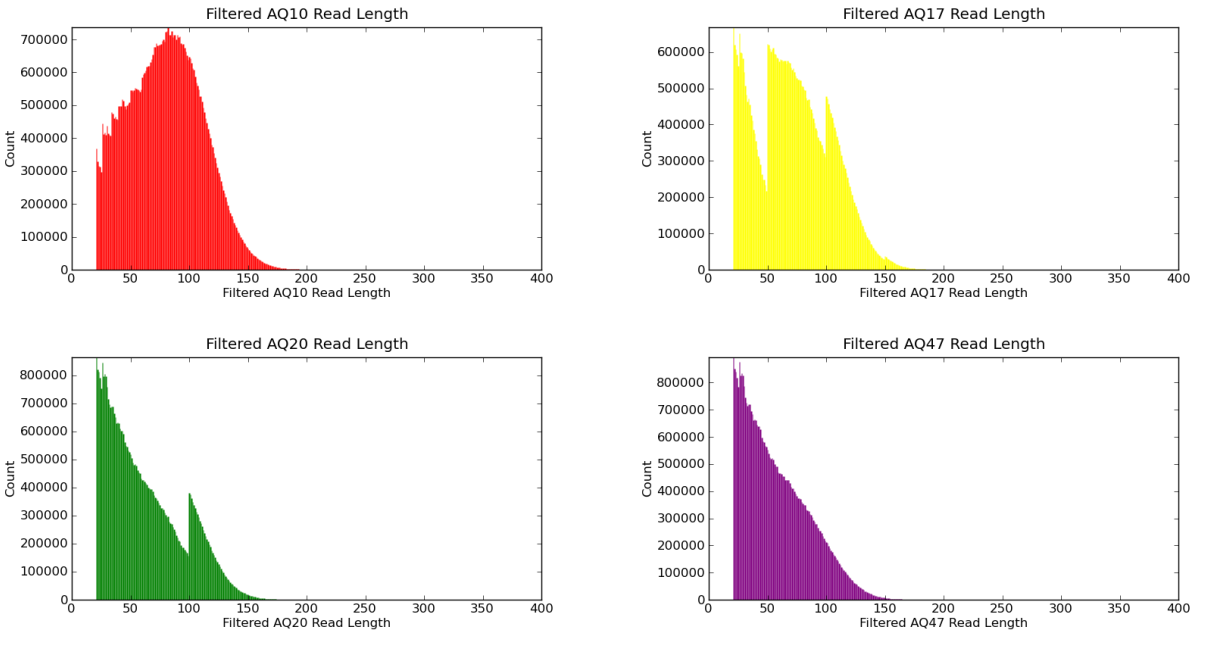
1.7X
Average Coverage Depth of Reference



97.2%
Mean Raw Accuracy 1x



	AQ17	AQ20	Perfect
Total Number of Bases [Mbp]	3.5 G	2.7 G	2.5 G
Mean Length [bp]	71	62	57
Longest Alignment [bp]	240	231	230
Mean Coverage Depth	1.1	0.9	0.8



variantCaller

Library type: Whole Genome
Targeted regions: None
Hotspot regions: None
Configuration: Germ Line - High Stringency (Proton)
Download all barcodes: [VCF.ZIP] [XLS.ZIP]

Barcode Name	Sample Name	Variants	Download Links
IonXpress_017	M-28294	38745	[VCF.GZ] [VCF.GZ.TBI] [XLS]
IonXpress_018	N-43592	48153	[VCF.GZ] [VCF.GZ.TBI] [XLS]
IonXpress_019	O-43364	40640	[VCF.GZ] [VCF.GZ.TBI] [XLS]
IonXpress_020	P-43614	36455	[VCF.GZ] [VCF.GZ.TBI] [XLS]
IonXpress_021	Q-43863	43434	[VCF.GZ] [VCF.GZ.TBI] [XLS]

Analysis Details

Run Name	R_2013_09_04_14_43_43_user_PRO-8-Methylation_MNOPQ
Run Date	Sept. 4, 2013, 2:44 p.m.
Run Flows	400
Projects	FX_Methylation
Sample	P-43614, Q-43863, M-28294, N-43592, O-43364
Reference	
PGM	Proton
Flow Order	TACGTACGTCTGAGCATCGATCGATGTACAGC
Library Key	TCAG
TF Key	ATCG
Chip Check	Passed
Chip Type	900
Chip Data	tiled
Barcode Set	IonXpress
Analysis Name	Auto_user_PRO-8-Methylation_MNOPQ_30
Analysis Date	Sept. 4, 2013, 11:04 p.m.
Analysis Flows	0
runID	NQU8Z

Software Version

Torrent_Suite	3.6.2
host	J7RBQV1
ion-alignment	3.6.3-1
ion-analysis	3.6.39-1
ion-dbreports	3.6.52-1
ion-gpu	3.6.5-1
ion-pipeline	3.6.25-1
ion-plugins	3.6.45-1
ion-protonupdates	3.6.7
ion-torrentr	3.6.9-1
LiveView	1612
DataCollect	2662
OIA	65
OS	16
Graphics	30